HAND DETECTION IN STATIC IMAGES, VIDEO SEQUENCES AND REAL TIME CAMERA FEED

Tomáš Bravenec

Master Degree Programme (2), FEEC BUT

E-mail: xbrave01@stud.feec.vutbr.cz

Supervised by: Tomáš Frýza

E-mail: fryza@feec.vutbr.cz

Abstract: The goal of this project is to create a computer vision system capable of hand detection in static images and in video sequence either from existing recording or real time feed from connected camera. Algorithms commonly used for hand detection are mostly dependent on simple background and are very dependent on the lightning changes. To mostly eliminate these issues this project uses deep convolutional neural network trained for hand detection.

Keywords: Computer Vision, Hand detection, Convolutional Neural Networks, Deep Learning

1 INTRODUCTION

As the computational power available to us in our everyday life steadily grows, the field of computer vision becomes more accessible and popular. Even mobile devices nowadays can easily detect human faces in pictures, or recognize the scene in front of the camera, so they can select appropriate camera setting for the best picture quality. One of the more difficult tasks is detection of an object that has a lot of different variations that unlike faces cannot be described with basic shape-based rules. One of these objects is human hand, that looks different depending on the angle of observation and just simple fist looks very different from open palm.

The ability to detect hands is helpful in more than one way. With the fact that we can track hand movement, it can be translated to easy hand gestures like moving of hand from left to right to switch playing tracks in car without looking for controls. Also, hand detection is just a first step to gesture recognition, which could help people suffering from vision loss understand more of what pose is in the person in a painting or what is happening in a video sequence. This could also help with automated sign language translation.

2 CONVOLUTIONAL NEURAL NETWORKS

Due to the fact, that rule-based detection can get confused easily just by changing lightning conditions or with more complicated background, more complex approach is needed. Because of this the neural networks were modeled after a human brain, just like brain neural network is composed of many “neurons” which pass information and do mathematical operations on inputted data.

The convolutional neural networks are mostly used in image processing domain. This type of network is created out of neurons, that in the training part, learn what kind of feature should they look for. By adding these layers, the neural network can learn more and more complex features and from these features decide on the output. Typical convolutional neural network is shown on Figure 1.

Creating neural networks from scratch is not impossible but training it from scratch can take a long time to get usable results. To avoid this and make the time needed to train the network as short as possible, it is possible to modify existing architecture and retrain it for detection of different objects, than it was trained to recognize in the first place. This approach is possible, because most neural networks for image processing follow the same pattern, find the most basic features like lines and

circles, and from this move to combining these patterns, which is followed up by prediction of result itself. This means that the first couple of layers can be reused and only the last few layers must be changed to recognize different patterns.

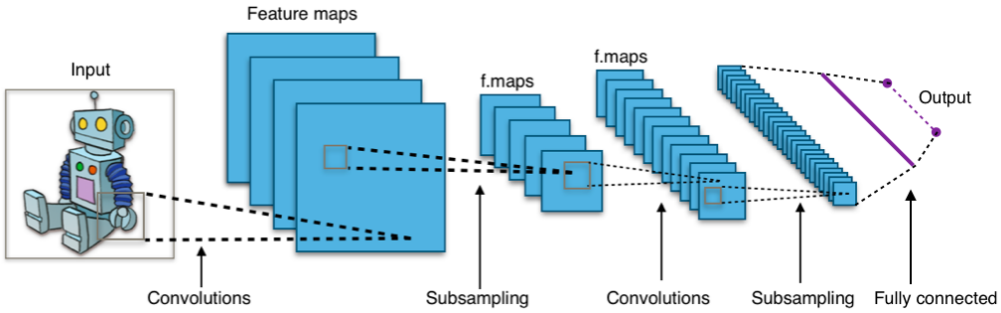

Figure 1: Typical convolutional neural network [1]

2.1 ARCHITECTURE YOLO V2

For the purpose of hand detection, the neural network of architecture YOLO v2 [2] was selected due to its high accuracy and very high speed. This architecture is designed in a way, that for detection of all objects in an entire image, unlike with most of other networks, is necessary only one forward pass through the network, from that goes the name You Only Look Once. This also means that this architecture is very well suited for real time processing.

By default, the network is designed for recognition of 80 different objects, this meant that the network had to be modified to recognize only single object – human hand. That was done in the configuration of the network by lowering the number of filters in the layer preceding the detection layer of the network in conformity with equation (1), from 255 to 30.

$$filters = (classes + 5) * 5 \quad (1)$$

YOLO v2 with pretrained weights comes in couple of versions which differ from each other with the input resolution of an image. The higher resolution, the more accurate but slower the network will be and vice versa. Higher resolution can be helpful when the network needs to detect high number of classes, but for the purpose of this project, the fastest network is the ideal choice.

2.2 TRAINING DATASETS

To get usable predictions out of a neural network, it needs a lot of data for training. For this was used a combination of EgoHands dataset [3], which contains images of people playing games from an egocentric point of view, images from New Zealand Sign Language dictionary **Chyba! Nenalezen zdroj odkazů.** and part of the MPII human pose estimation dataset [4]. This combination of training data allowed the network to generalize well for almost any situation. Examples of images from all three datasets are on Figure 2.

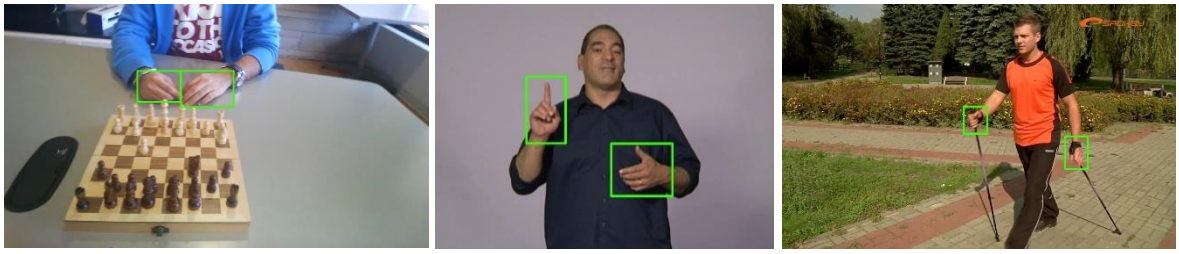

Figure 2: Images from used datasets with bounding boxes around hands

First thing before actual training, the annotations of the EgoHands dataset had to be converted to match the format of YOLO based architectures, and for sign language and MPII datasets, the annotations had to be created from scratch.

3 RESULTS

From the nature of neural network, it is recommended to use GPU instead of a CPU, as there will be massive performance drop, while using regular CPU. Due to this was for testing of the real-time processing used as a main computing unit NVIDIA GeForce GTX 1060 6GB. With this GPU the detection process run at average of 28 frames per second and most of the GPUs currently (March 2019), are more powerful than used GPU, which means that the framerate on even better GPU could go a lot higher. As the neural network has fixed input size of 416x416 pixels, this framerate is independent of resolution.

During testing with videos, camera and images with difficult background, obscured parts of hands with another object, the neural network made most of the predictions correctly. Neural networks predictions in couple of different conditions are shown on Figure 3.

Figure 3: Neural networks predictions on evaluation images and captured frame from camera

During automated evaluation over 847 images, from all used datasets the neural network managed to get to precision of 0.948, that means that 94.8% of all predictions were correct. Another important metric from evaluation is recall, which ended at value 0.702, which means that the neural network found 70.2% of all hands in evaluation part of dataset. This value is lower due to two hands overlapping each other, which network predicts to be one hand instead of two, this situation is not common in used datasets. To fix this, the network would have to be trained on larger datasets where this situation would occur more often.

4 CONCLUSION

Results of this project show, that the neural network generalized features of hands very well and can detect hands correctly with minimal error even in difficult conditions. Only issue is if the network is supposed to detect two overlapping hands, which results in only single detection. Solution to this issue would need including more of these situations in the training dataset and retraining the network with more data.

REFERENCES

[1] M. T. Islam, B. M. N. Karim Siddique, S. Rahman and T. Jabid: Image Recognition with Deep Learning. In *2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, October 2018.

[2] J. Redmon, A. Farhadi: YOLO9000: Better, Faster, Stronger, *arXiv preprint arXiv:1612.08242*, 2016.

[3] S. Bambach, S. Lee, D. J. Crandall and C. Yu: Lending A Hand: Detecting Hands and Recognizing Activities in Complex Egocentric Interactions. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[4] *New Zealand Sign Language Dictionary* [online]. Available: https://www.nzsl.nz/

[5] M. Andriluka, L. Pishchulin, P. Gehler and B. Schiele: 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.